Received: 27 March 2021 | Revised: 14 November 2021 | Accepted: 27 December 2021

IET Renewable Power Generation

DOI: 10.1049/rpg2.12394

ORIGINAL RESEARCH PAPER

Data-driven generation of synthetic wind speeds: A comparative study

Daniele D'Ambrosio[1] | **Johan Schoukens**[1,2] | **Tim De Troyer**[1] | **Miroslav Zivanovic**[3] | **Mark Charles Runacres**[1]

[1] Department of Engineering Technology (INDI), Vrije Universiteit Brussel (VUB), Brussels, Belgium

[2] Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands

[3] Electrical Engineering and Communication Department, Universidad Pública de Navarra, Pamplona, Spain

Correspondence
D. D'Ambrosio, Department of Engineering Technology (INDI), Vrije Universiteit Brussel (VUB), Pleinlaan 2, 1050 Brussels, Belgium.
Email: daniele.dambrosio@vub.be

Funding information
Fonds Wetenschappelijk Onderzoek, Grant/Award Number: 74213/K231719N

Abstract

The increasing sophistication of wind turbine design and control generates a need for high-quality wind data. The relatively limited set of available measured wind data may be extended with computer generated data, for example, to make reliable statistical studies of energy production and mechanical loads. Here, a data-driven model for the generation of surrogate wind speeds is compared with two state-of-the-art time series models that can capture the probability distribution and the autocorrelation of the target wind data. The proposed model, based on the phase-randomised Fourier transform, can generate wind speed time series that possess the power spectral density of the target data and converge to their generally non-Gaussian probability distribution with an arbitrary, user-defined precision. The model performance is benchmarked in terms of probability distribution, power spectral density, autocorrelation, and nonstationarities such as the diurnal and seasonal variations of the target data. Comparisons show that the proposed model can outperform the selected models in reproducing the statistical descriptors of the input datasets and is able to capture the nonstationary diurnal and seasonal variations of the wind speed.

1 | INTRODUCTION

In recent years, the penetration of wind power in the electricity systems has increased considerably [1]. As a result, a growing need to efficiently integrate the increasing share of wind energy into the grid has emerged [2]. The availability of high-quality wind speed data has become crucial to advance the integration process while keeping the cost of wind energy low [3]. However, due to the cost and duration of wind measurement campaigns it has become increasingly advantageous to rely on surrogate wind data for the development of several strategic applications. In particular, the latest advancements in power system modelling with an increased share of wind energy [4, 5], in the design of larger and lighter rotors as well as in control and condition monitoring strategies have created a need for realistic surrogate time series of wind speeds. The recent increase in the use of sonic anemometers as well as reanalysis data such as MERRA-2 [6] has brought the advent of high-quality datasets that can be used to develop and tune wind speed models for the generation of realistic wind speed time series.

The generation of surrogate time series is referred to in the meteorology and wind energy communities as wind speed modelling, that is related to, but not synonymous with, forecasting. The goal of wind speed modelling is not to predict the future as in forecasting, but to computer-generate surrogate data that share as many relevant features as possible with physical data. Wind time series are characterised by probability density functions (PDF), expressing the relative frequency of occurrence of wind speeds, and by power spectral densities (PSD) or, equivalently, autocorrelation functions (ACF), expressing the temporal coherence of the data. The PDF of wind speeds is typically non-Gaussian, unless very short timespans are considered. The positively skewed Weibull distribution is the most commonly used distribution for wind data [7]. The PSD characterises the wind time series in terms of the dominant frequencies and the related temporal patterns that drive the wind speed process [8]. Moreover, wind time series are inherently non-stationary as their PDF and PSD vary over time as a result of deterministic meteorological factors changing over diurnal and seasonal time scales [9–11]. Therefore, capturing all the above-mentioned features poses a major challenge for any wind speed model in the generation of realistic surrogate wind speed time series.

There exist different families of models to generate surrogate time series, more or less common depending on the branch of physics or engineering where they are applied. The first distinction to make is between physical modelling on the one hand and data-driven (statistical) modelling on the other. Physical modelling generally involves the solution of physical conservation equations, as is done in numerical weather prediction tools such as WRF [12]. Physical models are highly effective, but have a high computational cost and may produce more information than is required. Data-driven models on the other hand, do not involve the solution of physical conservation laws, although they may use physical constraints on the model output, and are comparatively cheap. All data-driven models start from a set of training data and learn from those data how to produce surrogate time series with the same characteristics as the training data.

In the data-driven or statistical wind modelling class, established methods for generating surrogate wind speeds broadly fall in one of the two categories: Markov chain (MC)-based models and autoregressive integrated moving average (ARIMA) models. Markov chain models do not require the estimation of a continuous wind speed distribution and as such can conform to a discrete measured distribution. The fact that they are capable of reproducing the PDF of the real data is indeed one of their main advantages [13]. Their main drawback is that their memory is limited by the order of the model. As a consequence, daily trends can only be modelled with time steps of a few hours. Equivalently, the 10 min time steps prescribed by the IEC 61400-12-1 standard [14] only allow very short-term trends to be modelled. There is a practical limit to expanding the time memory of Markov chain models by adding extra orders, as the models quickly become prohibitively expensive with increasing model order. Nesting Markov chain models can improve the performance over standard Markov chain models but even for those models the autocorrelation quality quickly deteriorates as the time lag becomes larger [15].

ARIMA models consist of a modified form of the autoregressive-moving average (ARMA) process, that are designed to model time series with a homogeneous non-stationary behaviour [16]. In its standard modelling procedure, it is not well suited to model highly non-stationary and non-Gaussian random processes as the wind speed variation and thus requires modifications to produce satisfactory simulation results [17]. A proper power transformation of the data can be introduced to partly overcome the limitation to Gaussian processes; with the further introduction of limitation and seasonal partition of the data, it has been shown to adequately model the monthly variation of wind power generation [18]. Furthermore, ARIMA models struggle to capture reliably the diurnal and seasonal variations of wind speeds. This can be partly mitigated by using nested ARIMA methods, but even then the distributions of the wind data are far from perfect [19]. Even with improvements such as nesting, neither the ARIMA nor the Markov chain models are fully satisfactory for wind modelling, as manifested in their inability to conform to both the PDF and the PSD (or the ACF) of a measured dataset.

A third and prominent category of data-driven models based on artificial neural networks (ANN) has been proposed for wind scenario generation and forecasting. Among those models, machine learning algorithms based on generative adversarial networks (GAN) have been shown to be capable of generating realistic wind power scenarios on a limited time horizon of a few days that conform simultaneously to the PDF and the PSD of the test data [20, 21]. However, to our knowledge no investigation has been conducted on the capability of these models of capturing temporal correlations and probability distributions over longer time horizons such as seasonal and annual variations.

Recently, an alternative data-driven method for the generation of synthetic wind speeds was suggested by the authors in reference [22], which is based on the non-Gaussian phase-randomised Fourier transform (NGPRFT) model. The class of NGPRFT data-driven models originated in the 1990s in the fields of non-linear physics [23, 24] and system identification [25], and is able to produce surrogate data that do conform to both a prescribed generally non-Gaussian PDF and a PSD. Prior to reference [22], this class of models was never considered for the generation of surrogate wind data. However, because of their iterative rank-reordering process, the NGPRFT model class is a promising alternative as it can also capture non-stationary features of the wind speed such as its diurnal and seasonal variations.

The main contribution of this work is to compare the NGPRFT model from reference [22] with recent implementations of the Markov chain [26] and ARIMA [17] models for the generation of synthetic wind data that aim to reproduce the diurnal and seasonal variations of the wind speed. The proposed model is applied to the same datasets used for the published test cases of the selected models, and the simulation results are compared in terms of the accuracy in reproducing the PDF, the PSD, and the non-stationary features of the input data. Additionally, the level of user interaction required by the selected models is discussed and compared with the NGPRFT model. A second contribution of this paper is to test and discuss the performance of the proposed NGPRFT model when wind speed datasets of different time resolutions and record lengths are used to generate surrogate data.

2 | PHASE-RANDOMISED FOURIER TRANSFORM MODEL

This section briefly describes the main implementation steps of the proposed NGPRFT methodology. The reader is referred to reference [22] for a thorough description of its algorithm. The proposed model is illustrated in the high-level flow chart of Figure 1. In the initialization phase, two initial sequences are generated that are consistent with the PSD and the PDF of the target wind speed time series, respectively. Next, an iterative process first reorders the sequence conforming to the target PDF to match the rank order of the sequence possessing the target PSD. Then, a new sequence is generated from the Fourier amplitudes of the target data and the spectral phases of the

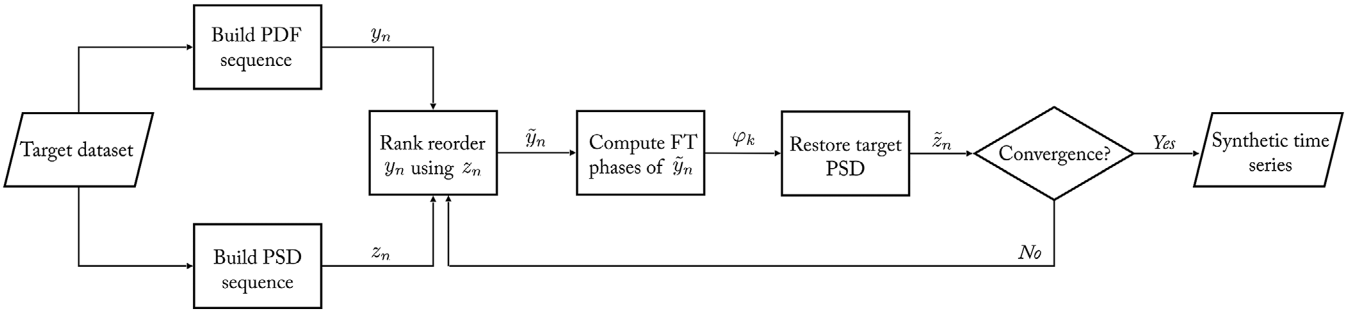

FIGURE 1 High-level flowchart of the proposed NGPRFT model

last reordered sequence. This new sequence replaces the initial sequence with the target PSD in the next reordering step, and the iterations continue until the last generated sequence has converged to the PDF of the target data.

2.1 | Initialization phase

2.1.1 | PSD sequence

The initial sequence possessing the target PSD is generated as a random-phase multi-sine signal from the frequency domain. This is achieved by inverse Fourier transforming the absolute Fourier amplitudes of the target data $|X(f_k)|$ multiplied by a set of phases φ_k^{rnd} randomly sampled from the uniform distribution $\mathcal{U}[0, 2\pi)$, where $k = 0, \ldots, N-1$ and $|X(f_k)| = 0$ for $k = N/2, \ldots, N-1$, with N equal to the total number of observations in the target time series; that yields

$$z_n = 2\Re\{\mathcal{F}^{-1}\{|X(f_k)|e^{j\varphi_k^{\mathrm{rnd}}}\}\}, \tag{1}$$

where $\Re\{\cdot\}$ indicates the real part, and $n = 0, \ldots, N-1$.

2.1.2 | PDF sequence

The initial sequence consistent with the PDF of the target wind speed dataset is obtained by performing an inverse cumulative distribution function (CDF) transform on an equally spaced sequence of N samples taken on the interval $(0, 1)$, and denoted by u_n,

$$y_n = F^{-1}(u_n), \tag{2}$$

where F^{-1} is the inverse CDF of the target wind speed time series. To complete the initialization phase, the variance and the mean of the sequence consistent with the target PDF, y_n, are imposed on z_n, and the absolute values of the Fourier amplitudes of this sequence are stored, $Z_k = |\mathcal{F}\{z_n\}|$.

2.2 | Iterative process

This process consists of three steps (Figure 1). First, a rank-reorder step is performed by applying a non-linear transformation that reorders y_n in a new sequence $\tilde{y}_n$; as a result, the smallest value of y_n is given the same position in $\tilde{y}_n$ that the smallest value of z_n has in its own sequence, and so forth for all the N values. In the second step, the spectral phases of this new sequence $\tilde{y}_n$ are computed by means of the Fourier transform of the signal, which yields

$$\varphi_k = \tan^{-1}\left(\frac{\Im\{\mathcal{F}\{\tilde{y}_n\}\}}{\Re\{\mathcal{F}\{\tilde{y}_n\}\}}\right), \tag{3}$$

where $\Im\{\cdot\}$ and $\Re\{\cdot\}$ indicate the imaginary part and the real part, respectively. In the last step, the final sequence $\tilde{z}_n$ is generated by inverse Fourier transforming the stored Fourier amplitudes Z_k multiplied by the last computed phases φ_k, and then taking the real part of the inverse Fourier transform; that is ,

$$\tilde{z}_n = \Re\{\mathcal{F}^{-1}\{Z_k e^{j\varphi_k}\}\}. \tag{4}$$

This last generated sequence replaces z_n in the successive rank-reordering step of y_n performed in the next iteration, and this three-step process is iterated until the PDF of $\tilde{z}_n$ converges to the PDF of the target wind speed dataset.

2.3 | Model properties

By its design, the proposed NGPRFT model simulates an ergodic, pseudo-random process. This emphasises two central features of the model. First, each realisation of the simulated process results in a synthetic time series that always conforms to the same statistical descriptors which are, by construction of the model, the PDF and the PSD of the input wind data. Second, the initial random phases φ_k^{rnd} give rise to a stochastic reordering of the synthetic wind speeds without altering their value in a random fashion, and that effectively delivers the same synthetic wind speeds with a different time evolution for each realisation of the model.

As for the extreme wind speeds, repeating the simulation with the same input data and with the same sampling defined by Equation (2) results in a different synthetic signal characterized by the same extreme values. These extremes, however, appear at different time instants in the synthetic signal as a result of the different random seed drawn for the initial phases φ_k^{rnd}. Therefore, if one wishes to extend the simulated extreme winds, the number of the simulated samples N is to be increased so as to increase the sampling in the tail region of the target CDF.

Periodic features of the target data such as the diurnal and seasonal variations create peaks in the PSD. The number of components in Equation (4) that is needed to model these periodic variations is minimized if an integer number of years (the seasonal variations) and days (the diurnal variations) are included in the synthesized signal. This leads to a sparse representation (only a few parameters are needed) of the periodic phenomena in the synthetic data that better mimics the observed seasonal/diurnal behaviour in the target data.

The random-phase multi-sine signal of Equation (1) requires that the total number of samples of the target data are used to retrieve the target PSD. Concurrently, the target PDF sequence of Equation (2) must possess an equal number of samples for the rank-reordering step to be enforced, which yields a synthetic sequence with the same number of samples as the target data. Therefore, the whole range of target data constitutes simultaneously the training and test dataset for the proposed model.

3 | COMPARISON WITH MC AND ARIMA MODELS

The performance of the proposed NGPRFT model is compared with two state-of-the-art modelling techniques for the simulation of non-stationary wind speeds, namely the non-homogeneous Markov chain model (NHMC) put forward by Xie et al. [26], and the ARIMA-based frequency-decomposed methodology presented by Yunus et al. [17]. These models are selected as a benchmark since both aim to reproduce the probability distribution and the time correlation of the observed data, while attempting to capture seasonal and diurnal variations that are characteristic of the wind speed stochastic process. Such non-stationary features are taken into account by adopting different modelling strategies.

The comparison carried out in this section aims at showing the performance of the proposed NGPRFT model in reproducing the PDF and the PSD, or equivalently, the ACF of the observed wind data when it is applied to the same datasets of the test cases presented in references [26] and [17] for the NHMC model and the ARIMA-based model, respectively. Particular emphasis is put on how accurately the proposed model can capture the non-stationarity of the wind speed data pertaining to the seasonal and diurnal variation of the wind speed. In addition, the level of user interaction required by the selected models is discussed.

3.1 | NHMC model comparison

In the NHMC model, the time homogeneity assumption is relaxed allowing the Markov chain transition probability matrix to become a function of time. Then, the time-varying transition matrix, that represents the wind speed variation at different times, is generated by means of a seasonal partition technique and a sequence period extraction procedure is performed on the wind data. This allows to yield state transition probabilities that are time related and finely adjusted to simulate the diurnal and seasonal variation of the wind speed.

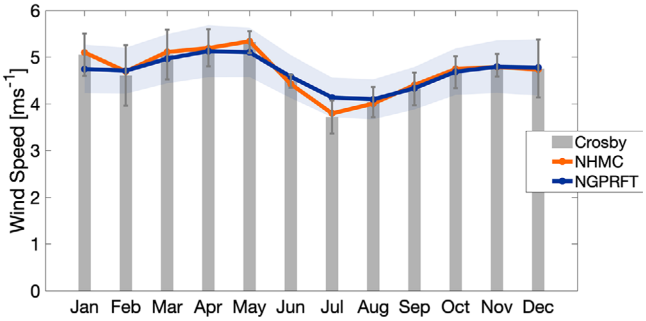

FIGURE 2 Average seasonal variation comparison shown as monthly average wind speeds of 10 years of data. ASV observed at Crosby as grey bars along with its inter-annual variability as grey error bars; NGPRFT-generated ASV as blue, thick line with its confidence intervals as blue, shaded regions (average over 10,000 simulations); NHMC-generated ASV as orange, thick line

The proposed NGPRFT method is applied to the same dataset used in the case study of reference [26], recorded at Crosby, USA, that is the hourly averaged wind speeds extracted from the database available on the internet at the North Dakota Agricultural Weather Network (NDAWN) website [27] for the 10-year period ranging from January 2003 to December 2012. A synthetic wind speed time series of the same length is generated, and the agreement of the proposed model with the target data is investigated in terms of its probability distribution, its PSD, and its ACF. To allow for a direct comparison, the same statistical descriptors produced by the NHMC model and shown in the case study of reference [26] are digitised and presented along with the results given by the NGPRFT model. In addition, the degree of non-stationarity reproduced in the synthetic data is shown and compared in terms of the average seasonal variation (ASV) of the wind speed and the amplitudes of the diurnal cycle harmonics of the wind speed detected in the PSD.

3.1.1 | Average seasonal variation

Figure 2 shows the ASV of the observed wind speeds at Crosby calculated as the monthly average wind speed across the 10 years of data (grey bars), along with its inter-annual variability calculated as the associated standard deviation (grey error bars). The same average variation is shown for the synthetic data simulated by the NGPRFT model (blue thick line) and by the NHMC model (orange thick line). The ASV shown for the NGPRFT model is an average result obtained over 10,000 realizations of the model; the confidence strips of the multiple realizations are calculated as the associated standard deviation (blue shaded regions), and indicate the inter-annual variability of the monthly average synthetic data. This is done as the initial random seed for the phases φ_k^{rnd} determines a stochastic rearrangement of the synthetic wind speeds that ultimately leads to an ASV varying in a limited range with the different realizations of the model.

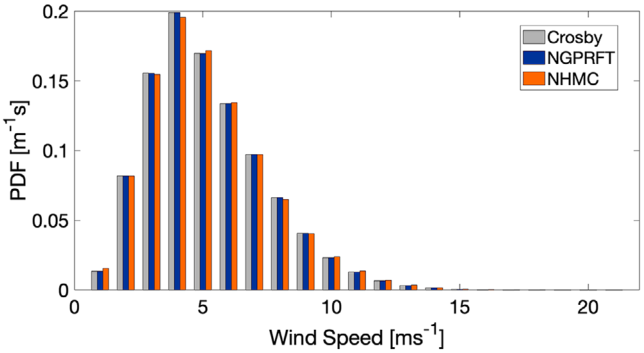

FIGURE 3 PDF agreement of synthetic wind speeds generated by the NGPRFT model and by the NHMC model with the wind data observed at Crosby

The monthly fluctuation observed in Figure 2 reveals a significant intra-annual or seasonal variation of the wind speed in the observed data at Crosby. This seasonal variation is well reproduced by the proposed NGPRFT model, whose ASV averaged across multiple simulations is consistent with the ASV of the observed wind data. From a visual comparison with the ASV produced by the NHMC model it can be noticed that overall the NGPRFT model shows a comparable performance with the model in reference [26] in simulating the observed seasonal variation at Crosby. For some months, the NGPRFT model yields slightly larger deviations from the observed data compared to the NHMC model (January, May, and July), showing a poorer performance. However, note that the NGPRFT-simulated ASV is presented here as an average over multiple realisations to show its convergence to the observed data; when simulating, the ASV error can be computed and one may reject simulation results until a prescribed tolerance is satisfied. In contrast, the authors of reference [26] do not specify whether their simulation result represents one realisation of the NHMC model or is an average over multiple realisations. Nevertheless, a degree of randomness in the simulation results of the NHMC model is to be expected as Markov chain methods belong to the class of Monte Carlo methods.

3.1.2 | Probability distribution

The PDF of the synthetic wind speeds generated by the proposed model is shown in Figure 3 along with the probability distributions of the observed wind speeds at Crosby and the synthetic speeds produced by the NHMC model. It can be noticed that the PDF yielded by the NGPRFT model is consistent with the observed wind speed distribution, and its simulated wind speeds fit very accurately the target stochastic process. The goodness of the fit with the observed wind data is evaluated in terms of the R^2 coefficient and the root-mean-square error (RMSE) of the CDF, RMSE_{CDF}. In Table 1, these statistics are summarised and compared with the values obtained from the application of the NHMC model and provided in reference [26]. The visual comparison and the reported metrics show that the NGPRFT model and the NHMC model attain a similar performance in reproducing the probability distribution of the wind speed measured at Crosby.

TABLE 1 Goodness of fit of the compared models for the Crosby dataset

Site	Model	R^2	RMSE_{CDF}
Crosby			
	NGPRFT	0.999999	0.0005
	NHMC	0.999963	0.0020

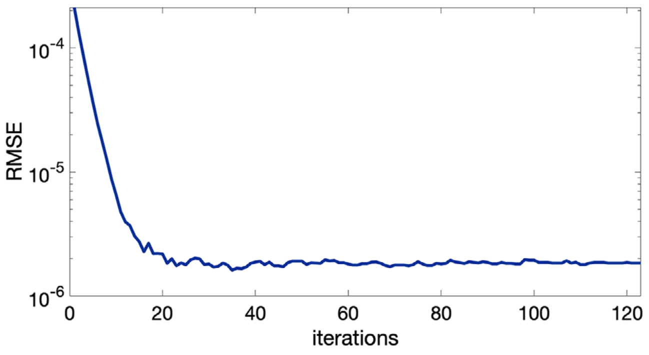

FIGURE 4 Root-mean-square error of the NGPRFT-generated PDF with respect to the PDF of the Crosby wind data

TABLE 2 ACF error produced by the compared models for different lags when synthesizing from the Crosby dataset

Model	12-h lag	24-h lag	48-h lag	100-h lag
NGPRFT	0.0001	0.0001	0.0001	0.0001
NHMC	0.0247	0.0193	0.0160	0.0157

It is important to note that the observed PDF is always reproduced with the same level of accuracy when multiple simulations are carried out with the proposed NGPRFT model. This means that the random phases drawn to generate the initial PSD sequence (random-phase multi-sine) do not affect the values of the generated wind speeds but only determine their reordering in the time series.

A measure of the convergence rate of the NGPRFT model is given in Figure 4, where the deviation of the synthetic PDF from the target PDF is calculated at each iteration as the RMSE. A fast convergence can be observed.

3.1.3 | Autocorrelation function

The analysis of the ACF of the wind speed modelled by the proposed NGPRFT model resulted in the agreement shown in Figure 5. The synthetic ACF is presented for the first 48 lags, corresponding to 48 h, along with the observed ACF at Crosby, and it is compared with the ACF produced by the NHMC model for the same number of lags. In addition, the RMSE of the ACF, RMSE_{ACF}, is calculated for four selected time lags and compared in Table 2 with the same ACF error produced by the NHMC model and presented in the case study of reference [26].

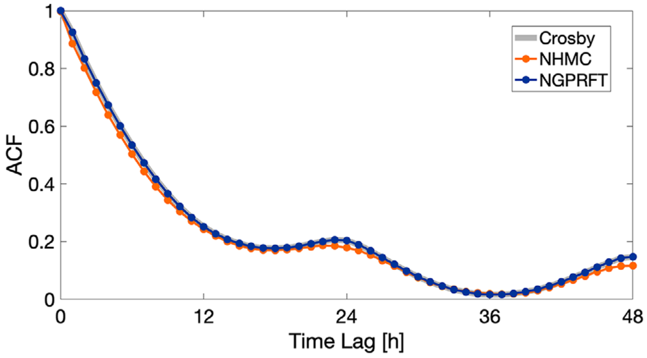

FIGURE 5 ACF agreement of synthetic wind speeds obtained from the NGPRFT model and from the NHMC model with wind data measured at Crosby

This analysis shows that the proposed NGPRFT model generates synthetic wind speeds whose ACF is consistent with the ACF of the wind speed observed at Crosby. Figure 5 illustrates that the proposed model can reproduce accurately the diurnal correlation of the observed wind speed indicated by the periodic variation of the ACF with a 24-h period. Also, a visual comparison reveals that the NHMC model yields a similar agreement with the observed ACF, being able to capture both the ACF decaying trend and the diurnal correlation structure of the wind speed at Crosby. From the quantitative comparison presented in Table 2, it can be noticed that the NGPRFT model yields an RMSE in the ACF that is consistently lower than the ACF error produced by the NHMC model for all the analysed time lags.

When multiple simulations are performed with the NGPRFT model, the target ACF is always reproduced with the same level of accuracy as the initial random phases do not affect the autocorrelation structure of the simulated process.

3.1.4 | Power spectral density

The performance of the proposed NGPRFT model in reproducing the spectral energy content of the observed wind speeds at Crosby is investigated through a spectral analysis of the simulated wind speeds. This analysis allows to reveal the presence of deterministic frequency components that give rise to nonstationarity in the wind speed data, and to assess how accurately these components are simulated by the synthetic data. To this end, the PSD of the NGPRFT-simulated wind speeds is shown in the top plot of Figure 6, along with the PSD of the observed wind speeds at Crosby. To allow for a direct comparison with simulation results given in reference [26], the PSD produced by the NHMC model is digitised and presented in the bottom plot of the same figure. The simulation results of reference [26] are presented in a separate plot as the PSD of the observed data shown there does not coincide with the PSD calculated from the observed wind speeds and shown in the top plot of Figure 6. This suggests that some technique was applied to reduce the variance of the estimated PSD of the observed data, and the same technique is likely to have been applied to the NHMC-simulated PSD. As the digitised PSD is given in m s$^{-1/2}$, the PSD obtained from the NGPRFT model is shown in the same units by taking the square root of its values. For visualisation purposes, the zero frequency or DC component of the PSD is not shown. The lowest spectral peak is the annual frequency component occurring at f = 3.171×10^{-8} Hz.

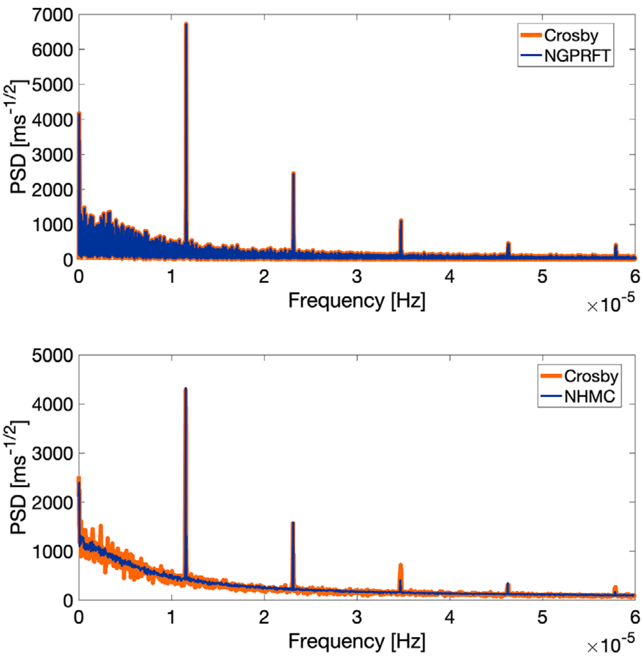

FIGURE 6 Square-root PSD agreement of synthesized wind speeds generated by the NGPRFT model (*top*) and by the NHMC model (*bottom*) with observed wind data at Crosby

The PSD analysis shows that the NGPRFT model can reproduce with high accuracy the entire frequency content of the wind speed at Crosby, including the strongest deterministic components detected in the observed data at the annual, diurnal, and semi-diurnal time scales, namely at f = 3.171×10^{-8}, 1.157×10^{-5}, and 2.314×10^{-5} Hz, respectively. The goodness of the agreement with the observed PSD is calculated as the RMSE in the PSD produced by the NGPRFT model, RMSE$_{\text{PSD}}$, that results in a value of 2.1×10^{-4}. The same level of accuracy in reproducing the observed PSD is attained for multiple realizations of the NGPRFT model, as the stochastic reordering due to the initial random phases φ_k^{rnd} always yields a synthetic wind speed time series consistent with the target PSD. Note that the presence of a strong diurnal component in the frequency domain is associated with the 24-h periodic variation observed for the ACF, as the power spectrum and the autocorrelation function constitute a Fourier-transform pair according to the Wiener–Khinchin theorem [28].

A visual inspection of the PSD agreement produced by the NHMC model (bottom plot of Figure 6) reveals that the NHMC-simulated PSD fails to reproduce the spectral content of the Crosby data with the same level of accuracy yielded by the NGPRFT model. The NHMC model underperforms in estimating the amplitudes of the observed PSD at Crosby throughout the analysed frequency range, with the exception of the diurnal and semi-diurnal harmonics whose amplitudes are correctly captured by the model. Larger deviations are observed for the amplitudes of the third and the fifth harmonics of the diurnal cycle. The poor retrieval of those harmonics suggests that the shape of the target diurnal cycle in the time domain is not reproduced with the same accuracy attained by the NGPRFT model.

In addition, the NHMC model fails to adequately reproduce the frequency content between the annual and the diurnal peak, that represents the wind speed fluctuations associated with the passage of large, synoptic-scale pressure systems [29]. As the authors of reference [26] do not provide any metric for the deviations from the observed PSD, it is not possible to perform a quantitative comparison with the error in the PSD obtained from the NGPRFT model.

3.1.5 | User interaction

As described in reference [26], the NHMC model construction requires user tuning during the preprocessing phase in order to enable the modelling of the seasonal and diurnal characteristics of the specific input wind data. In the seasonal effect partition, an optimal partition method is implemented to split the wind data in a number of segments that reflect the seasonal variability of the data. In this step, a user choice is required to determine the optimal number of segments as this is not specified by the optimal partition method. Then, after performing the sequence period extraction, the user has to decide on the most suitable wind variation period R according to the periodic characteristics of the wind data; this parameter will determine in turn the number of transition probability matrices used by the model to represent the wind speed variation at different times. Finally, any Markov chain model requires an initial choice by the user on the number of states into which the input wind data are discretised, that define the state transition probabilities of the transition matrix. Overall, the user interaction required during the preprocessing phase affects the performance of the NHMC model and necessitates some expertise by the user to fine tune the NHMC model.

In contrast, the NGPRFT model is fully automated and no tuning of the model is required to generate synthetic data from target data with any length and temporal resolution. It is only recommended that the target data contain an integer number of days and years so as to avoid introducing any leakage in the periodic components of the PSD.

3.2 | ARIMA model comparison

With regard to the ARIMA-based model, a frequency decomposition of the wind speed data is introduced in the standard ARIMA modelling procedure [16] to better capture the periodic and non-stationary characteristics of the wind speed fluctuations. This decomposition allows to split the wind speed data into a high-frequency (HF), stationary component, and a low-frequency (LF), non-stationary component that accounts for the seasonal and diurnal cyclical variation of the wind speed. Both components are in turn modelled separately by performing a standard ARIMA procedure, and then combined to get the synthetic wind speed time series. In addition, shifting and limitation of the wind speed data are introduced before modelling and during simulation with respect to the standard ARIMA modelling procedure.

The ARIMA-based frequency-decomposed model put forward in reference [17] is the second model selected for benchmarking the proposed methodology. To do that, the NGPRFT model is applied to generate synthetic wind data from the same dataset used in their test case, namely the 10-min average wind speeds recorded by the meteorological mast located in the Näsudden peninsula in Gotland, Sweden, from the 1 January to 31 December 2005 at a height of 100 m. This dataset is maintained and provided by the Department of Earth Sciences at Uppsala University. A synthetic wind speed time series of the same length of the dataset is generated, and the comparison with the ARIMA-simulated wind data is carried out in terms of the same statistical descriptors presented in Section IV of reference [17], namely the probability distribution, the ACF, and the power spectrum. Their values are digitised and shown along with the results given by the application of the proposed NGPRFT model.

3.2.1 | Probability distribution

The performance of the two models is first compared in terms of their capability to reproduce the probability distribution of the observed wind speeds. A quantile–quantile (Q–Q) plot is employed to assess such a capability. This type of plot provides a graphical method to compare two PDFs: the closer the data points lay on the straight line $y = x$, the better is the agreement of the compared probability distributions.

Figure 7 shows the Q–Q plot of the synthetic wind speeds obtained from the application of the proposed NGPRFT model, along with the digitised Q–Q plot produced by the ARIMA-based modelling. It can be noticed that the wind data simulated with the NGPRFT model are in very good agreement with the observed wind speeds at Näsudden. In contrast, the Q–Q plot given by the ARIMA-simulated data reveals deviations from the reference data occurring at wind speeds around 15 m s^{-1} that become larger with increasing wind speed. Additionally, significant deviations of the ARIMA-simulated data from the observed data can be also observed in the very low wind speed region between 0 and 3 m s^{-1}. The authors of reference [17] comment that the underperformance of the ARIMA model in reproducing the largest wind speeds is due to the limited number of wind speeds higher than 20 m s^{-1} in the Näsudden dataset. They further comment that the extreme wind conditions can be properly modelled with a separate technique when required.

In contrast, the extreme wind speeds observed at Näsudden are well captured by the NGPRFT model, that shows a very limited deviation with respect to the observed data for the highest wind speed occurrences. Moreover, the kernel of the proposed

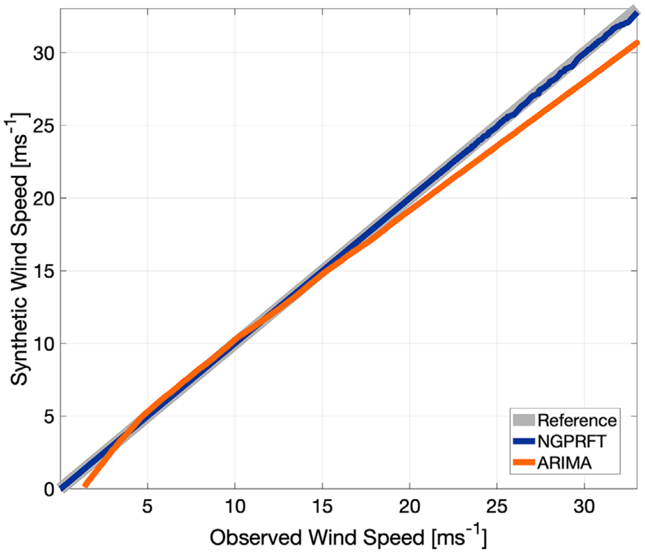

FIGURE 7 Quantile–quantile plot of observed wind speeds at Näsudden against synthetic wind speeds generated by the NGPRFT model and the ARIMA-based model.

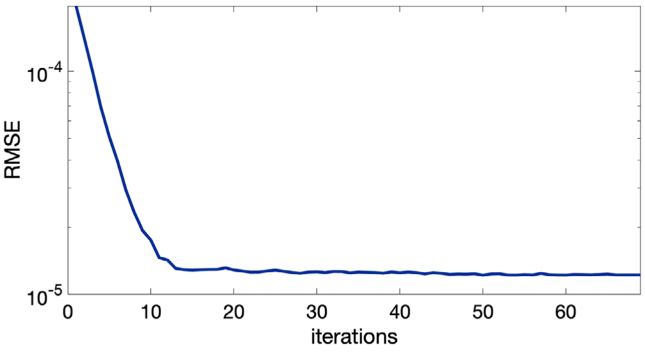

FIGURE 8 Root-mean-square error of the NGPRFT-generated PDF with respect to the Näsudden PDF.

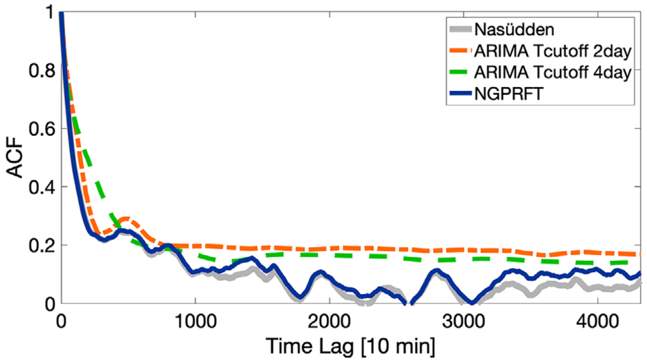

FIGURE 9 ACF agreement of the synthetic wind speeds simulated by the NGPRFT model and by the ARIMA-based model with the observed wind speeds at Näsudden.

model guarantees that the same extreme values are generated for each realisation of the NGPRFT model when it is applied to the same input data and with the same sampling defined by Equation (2). The stochastic component of the model also ensures that, for each simulation, the generated extreme wind speeds appear at different time instants in the synthetic time series as a result of the different random initial phases φ_k^{rnd}.

In addition, a measure of the convergence rate of the NGPRFT model is given in Figure 8, where the deviation of the synthetic PDF from the target PDF is calculated at each iteration as the RMSE. A fast convergence can be observed.

3.2.2 | Autocorrelation function

A second level of comparison is conducted to assess the performance of the two models, NGPRFT and ARIMA based, in simulating the temporal autocorrelation of the observed wind data at Näsudden. To do that, the ACF of the observed wind data is shown in Figure 9 along with the synthetic wind data generated by the NGPRFT model, and the digitised ACF values of the ARIMA-simulated data for a time lag up to one month (4320 lags). For the latter, the figure shows the ACF values resulting from two different cutoff frequencies ($1/T_{\mathrm{cutoff}}$), that determine different decompositions into low-frequency and high-frequency components in the ARIMA-based model.

A visual inspection reveals that the NGPRFT-simulated ACF follows very closely the autocorrelation of the wind speed measured at Näsudden throughout the whole range of analysed time lags. On the other hand, the ARIMA-simulated ACF that seems to yield a satisfactory match with the target autocorrelation profile ($T_{\mathrm{cutoff}} = 2$ days) can only reproduce the pattern of the observed ACF for approximately 500 lags, as its agreement degrades considerably at larger time lags. In addition, the ACF simulated by the ARIMA model shows a consistent bias throughout the calculated time lags. The authors of reference [17] deem the ACF agreement given by their ARIMA-based model satisfactory, and comment that the agreement after 500 lags is less significant as the ACF of the observed wind speed is lower than 0.2. However, this comparison shows that the proposed NGPRFT model performs significantly better in reproducing the observed ACF at Näsudden, even for values lower than 0.2.

3.2.3 | Power spectral density

Further information on how accurately the temporal autocorrelation and the periodic characteristics of the observed data are reproduced in the synthetic data can be inferred by performing a Fourier or spectral analysis on the observed and synthetised time series. For this reason, the periodograms of the measured and the simulated wind speed data is presented in reference [17]. To provide a meaningful comparison with their investigation, the periodograms of the observed data and the synthetic data obtained from the NGPRFT model are calculated and shown in the top plot of Figure 10 in the same units as in reference [17], namely dB (cycles/h)$^{-1}$. The bottom plot of Figure 10 shows the digitised ARIMA-simulated periodogram along with

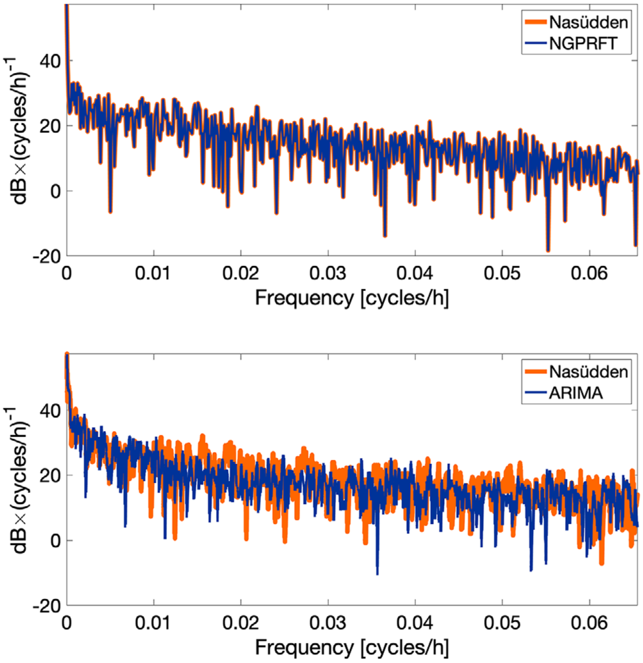

FIGURE 10 Periodogram agreement of synthetic wind speeds with Näsudden dataset. NGPRFT-simulated periodogram in the top plot; ARIMA-simulated periodogram in the bottom plot

the periodogram of the observed data calculated in reference [17].

A visual comparison of the top plot of Figure 10 reveals that the NGPRFT model can reproduce with very high accuracy the PSD of the observed wind speed time series throughout the whole range of computed frequencies. The deviation of the NGPRFT-simulated periodogram with respect to the observed periodogram is given as the RMSE of the statistics, RMSE_{PER}, which yields a value of 7.7×10^{-3}.

In contrast, the bottom plot of Figure 10 shows that the ARIMA-based model manages to capture only the decaying trend of the observed spectral content without matching accurately the magnitudes of the target periodogram. The authors of reference [17] do not provide a metric that quantifies the observed deviation, therefore only a visual comparison is possible. Nevertheless, the results shown in this second level of comparison suffice to state that the proposed NGPRFT model outperforms the ARIMA-based model in reproducing both the temporal autocorrelation and the spectral content of the observed wind data at Näsudden.

3.2.4 | User interaction

In the modified ARIMA-based modelling procedure proposed in reference [17], user interaction is required throughout the process. During the first stage, the HF and LF components are obtained by performing a standard ARIMA modelling procedure. This entails user intervention to determine the proper combination of required transformations (i.e. differencing and power transformation) to be applied to the observed wind data in order to identify the correct ARIMA model structure for the two components. In addition, a choice has to be made by the user for a suitable criterion to use for the determination of the transformation factor ν that yields the best simulation results for each frequency component. Although the subsequent steps of the model identification can be automated, the modified ARIMA-based model also introduces shifting and limitation of the observed wind data before modelling to improve simulation results. The necessity to implement both steps is left to the user to judge based on the characteristics of the input time series. In a positive case, user interaction is required to determine suitable upper and lower limits and/or a constant offset value to apply to the input data that can be estimated by performing sensitivity analysis on the simulation results. Overall, the fine tuning of the modified ARIMA-based model entails a high level of user interaction, that requires an expert time-series analyst with previous experience in ARIMA model identification to be performed effectively.

In contrast, the NGPRFT model does not require user interaction throughout its operation and thus can be fully automated. In particular, no tuning of the model is needed to generate synthetic wind speeds from target data of different length and temporal resolution.

4 | CONCLUSIONS

Here, the proposed NGPRFT model has been compared with two state-of-the-art models for the generation of surrogate wind data to benchmark its performance in reproducing the probability distribution, the PSD, and the periodic variations of the target wind speed dataset. The main contribution of this work has been to show that the NGPRFT class of data-driven models can outperform Markov chain and ARIMA models in generating surrogate data that conform to both the generally non-Gaussian PDF and the PSD of a given wind speed dataset. In addition, its performance in capturing the target diurnal and seasonal variations of the wind speed has been analysed.

The models selected for the comparison were the NHMC model of reference [26] and the ARIMA-based model in reference [17]. The NGPRFT model has been applied to the same datasets used in the respective test cases of the selected models, and the comparison has been conducted in terms of the PDF, the ACF, and the PSD of the generated time series. For both test cases, the NGPRFT-simulated wind speeds show a perfect reconstruction of both the PDF and the PSD of the target wind data. In terms of probability distribution, the NGPRFT model produces a marginally superior agreement with the target PDF compared to the NHMC model, whereas it yields a significantly better performance with respect to the ARIMA-based model. As for the PSD, the proposed model outperforms both the NHMC approach and the ARIMA-based model in reproducing the target PSD in the respective test cases. In particular, the NGPRFT-simulated PSD reproduces with high fidelity all the harmonics of the diurnal cycle in the test case of reference [26], while the NHMC-simulated PSD fails in getting the correct spectral amplitudes for some of those harmonics (third and fifth). The ACF analysis confirms this better performance and shows that the NGPRFT-simulated wind speeds reproduce the same diurnal correlation of the observed wind data. A substantially superior performance is also shown by the NGPRFT model in the ACF analysis of the test case of reference [17] compared to the ARIMA-based model. Additionally, the proposed model sufficiently reproduces the seasonal variations of the wind data in the test case of reference [26], showing the ability to capture such a non-stationary feature of the wind variation.

In addition, a user-interaction analysis shows that both the NHMC model and the ARIMA-based approach require user intervention to fine tune the modelling process according to the characteristics of the input wind data. In contrast, the proposed NGPRFT model does not require any tuning from the user and shows identical performance when applied to datasets with different temporal resolutions (1 h and 10 min, respectively), and different record lengths (10 years and 1 year, respectively).

ACKNOWLEDGEMENTS

This work was partly supported by the Research Foundation Flanders (FWO) [grant number 74213/K231719N].

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Daniele D'Ambrosio https://orcid.org/0000-0001-9331-3281
Johan Schoukens https://orcid.org/0000-0003-0492-6137
Tim De Troyer https://orcid.org/0000-0002-6598-7374
Miroslav Zivanovic https://orcid.org/0000-0001-8729-6657
Mark Charles Runacres https://orcid.org/0000-0002-8123-7637

REFERENCES

1. Lee, J., Zhao, F., Dutton, A., Backwell, B., Fiestas, R., Qiao, L., et al.: Global wind report 2019. Brussels, Global Wind Energy Council (GWEC) (2020)
2. Draxl, C., Clifton, A., Hodge, B.M., McCaa, J.: The wind integration national dataset (wind) toolkit. Appl. Energy. 151, 355–366 (2015). http://www.sciencedirect.com/science/article/pii/S0306261915004237
3. van Kuik, G.A.M., Peinke, J., Nijssen, R., Lekou, D., Mann, J., Sørensen, J.N., et al.: Long-term research challenges in wind energy - a research agenda by the european academy of wind energy. Wind Energy Sci. 1(1), 1–39 (2016). https://wes.copernicus.org/articles/1/1/2016/
4. Carapellucci, R., Giordano, L.: The effect of diurnal profile and seasonal wind regime on sizing grid-connected and off-grid wind power plants. Appl. Energy. 107, 364–376 (2013). http://www.sciencedirect.com/science/article/pii/S0306261913001529
5. Suomalainen, K., Silva, C.A., Ferrão, P., Connors, S.: Synthetic wind speed scenarios including diurnal effects. Implications for wind power dimensioning. Energy 37(1), 41–50 (2012). http://www.sciencedirect.com/science/article/pii/S0360544211005317
6. Gelaro, R., McCarty, W., Suárez, M.J., Todling, R., Molod, A., Takacs, L., et al.: The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). J. Clim. 30(14), 5419–5454 (2017). https://journals.ametsoc.org/view/journals/clim/30/14/jcli-d-16-0758.1.xml
7. Carta, J.A., Ramírez, P., Velázquez, S.: A review of wind speed probability distributions used in wind energy analysis: Case studies in the Canary Islands. Renewable Sustainable Energy Rev. 13(5), 933–955 (2009). http://www.sciencedirect.com/science/article/pii/S1364032108000889
8. Harris, R.I.: The macrometeorological spectrum–a preliminary study. J. Wind Eng. Ind. Aerodyn. 96(12), 2294–2307 (2008). http://www.sciencedirect.com/science/article/pii/S0167610508001025
9. Barthelmie, R.J., Grisogono, B., Pryor, S.C.: Observations and simulations of diurnal cycles of near-surface wind speeds over land and sea. J. Geophys. Res. Atmos. 101(D16), 21327–21337 (1996). https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/96JD01520
10. Dai, A., Deser, C.: Diurnal and semidiurnal variations in global surface wind and divergence fields. J. Geophys. Res. Atmos. 104(D24), 31109–31125 (1999). https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/1999JD900927
11. He, Y., Monahan, A.H., McFarlane, N.A.: Diurnal variations of land surface wind speed probability distributions under clear-sky and low-cloud conditions. Geophys. Res. Lett. 40(12), 3308–3314 (2013). http://doi.wiley.com/10.1002/grl.50575
12. Skamarock, C., Klemp, B., Dudhia, J., Gill, O., Barker, D., Duda, G., et al.: A description of the advanced research WRF version 3 (No. NCAR/TN-475+STR). (2008). https://opensky.ucar.edu/islandora/object/technotes%3A500/
13. Brokish, K., Kirtley, J.: Pitfalls of modeling wind power using Markov chains. In: 2009 IEEE/PES Power Systems Conference and Exposition, pp. 1–6 (2009)
14. IEC 61400-12-1:2017. Wind energy generation systems - Part 12-1: Power performance measurements of electricity producing wind turbines. Tech. rep. International Electrotechnical Commission (2017)
15. Tagliaferri, F., Hayes, B.P., Viola, I.M., Djokić, S.Z.: Wind modelling with nested Markov chains. J. Wind Eng. Ind. Aerodyn. 157, 118–124 (2016). http://www.sciencedirect.com/science/article/pii/S0167610515301720
16. Box, G.E.P., Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: Time series analysis: Forecasting and control. Wiley Series in Probability and Statistics. Wiley, New York (2015)
17. Yunus, K., Thiringer, T., Chen, P.: ARIMA-based frequency-decomposed modeling of wind speed time series. IEEE Trans. Power Syst. 31(4), 2546–2556 (2016)
18. Chen, P., Pedersen, T., Bak-Jensen, B., Chen, Z.: ARIMA-based time series model of stochastic wind power generation. IEEE Trans. Power Syst. 25(2), 667–676 (2010)
19. Sim, S.K., Maass, P., Lind, P.G.: Wind speed modeling by nested ARIMA processes. Energies 12(1), 69 (2019). https://www.mdpi.com/1996-1073/12/1/69
20. Chen, Y, Wang, Y., Kirschen, D., Zhang, B.: Model-free renewable scenario generation using generative adversarial networks. IEEE Trans. Power Syst. 33(3), 3265–3275 (2018)
21. Jiang, C., Mao, Y., Chai, Y., Yu, M., Tao, S.: Scenario generation for wind power using improved generative adversarial networks. IEEE Access 6, 62193–62203 (2018)
22. D'Ambrosio, D., Schoukens, J., Troyer, T.D., Zivanovic, M., Runacres, M.C.: Synthetic wind speed generation for the simulation of realistic diurnal cycles. J. Phys. Conf. Ser. 1618, 062019 (2020). https://doi.org/10.1088/1742-6596/1618/6/062019
23. Theiler, J., Eubank, S., Longtin, A., Galdrikian, B., Farmer, J.D.: Testing for nonlinearity in time series: The method of surrogate data. Physica D 58(1), 77–94 (1992). http://www.sciencedirect.com/science/article/pii/016727899290102S
24. Schreiber, T., Schmitz, A.: Improved surrogate data for nonlinearity tests. Phys. Rev. Lett. 77(4), 635–638 (1996). https://link.aps.org/doi/10.1103/PhysRevLett.77.635
25. Schoukens, J., Dobrowiecki, T.: Design of broadband excitation signals with a user imposed power spectrum and amplitude distribution. In: IMTC/98 Conference Proceedings. IEEE Instrumentation and Measurement Technology Conference. Where instrumentation is going (Cat. No.98CH36222) 2, pp. 1002–1005 (1998)
26. Xie, K., Liao, Q., Tai, H., Hu, B.: Non-homogeneous Markov wind speed time series model considering daily and seasonal variation characteristics. IEEE Trans. Sustainable Energy 8(3), 1281–1290 (2017)
27. North Dakota Agricultural Weather Network (NDAWN). https://ndawn.ndsu.nodak.edu. Accessed 30 Oct 2020

28. Wiener, N.: Time series. MIT press, Cambridge (1949)
29. Van der Hoven, I.: Power spectrum of horizontal wind speed in the frequency range from 0.0007 to 900 cycles per hour. J. Meteor. 14(2), 160–164 (1957). https://journals.ametsoc.org/doi/abs/10.1175/1520-0469%281957%29014%3C0160%3APSOHWS%3E2.0.CO%3B2

How to cite this article: D'Ambrosio, D., Schoukens, J., De Troyer, T., Zivanovic, M., Runacres, M.C.: Data-driven generation of synthetic wind speeds: A comparative study. IET Renew. Power Gener. 16, 922–932 (2022). https://doi.org/10.1049/rpg2.12394